대한민국특허청
KOREAN INTELLECTUAL PROPERTY OFFICE

별첨 사본은 아래 출원의 원본과 동일함을 증명함.

This is to certify that the following application annexed hereto is a true copy from the records of the Korean Intellectual Property Office.

출 원 번 호 : 10-2003-0074429
Application Number

출 원 년 월 일 : 2003년 10월 23일
Date of Application OCT 23, 2003

출 원 인 : 한국전자통신연구원
Applicant(s) Electronics and Telecommunications Research Inst

2003 년 12 월 02 일

특 허 청
COMMISSIONER

【서지사항】

【서류명】	특허출원서
【권리구분】	특허
【수신처】	특허청장
【참조번호】	0002
【제출일자】	2003.10.23
【발명의 명칭】	유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치 및 그 방법
【발명의 영문명칭】	APPARATUS AND METHOD FOR RECOGNIZING BIOLOGICAL NAMED ENTITY FROM BIOLOGICAL LITERATURE BASED ON UMLS
【출원인】	
【명칭】	한국전자통신연구원
【출원인코드】	3-1998-007763-8
【대리인】	
【성명】	권태복
【대리인코드】	9-2001-000347-1
【포괄위임등록번호】	2001-057650-1
【대리인】	
【성명】	이화익
【대리인코드】	9-1998-000417-9
【포괄위임등록번호】	1999-021997-1
【발명자】	
【성명의 국문표기】	박수준
【성명의 영문표기】	PARK,Soo Jun
【주민등록번호】	660112-1052431
【우편번호】	135-110
【주소】	서울특별시 강남구 압구정동 477 현대아파트 203동 501호
【국적】	KR
【발명자】	
【성명의 국문표기】	김태현
【성명의 영문표기】	KIM,Tae Hyun
【주민등록번호】	760423-2951526

【우편번호】 306-020
【주소】 대전광역시 대덕구 대화동 16-178
【국적】 KR
【발명자】
【성명의 국문표기】 이현숙
【성명의 영문표기】 LEE,Hyun-Sook
【주민등록번호】 760303-2448921
【우편번호】 305-311
【주소】 대전광역시 유성구 구암동 432-4
【국적】 KR
【발명자】
【성명의 국문표기】 장현철
【성명의 영문표기】 JANG,Hyun Chul
【주민등록번호】 710920-1396542
【우편번호】 302-792
【주소】 대전광역시 서구 월평3동 황실아파트 107동 803호
【국적】 KR
【발명자】
【성명의 국문표기】 박선희
【성명의 영문표기】 PARK,Seon Hee
【주민등록번호】 580117-2069619
【우편번호】 302-741
【주소】 대전광역시 서구 만년동 강변아파트 112동 106호
【국적】 KR
【심사청구】 청구
【취지】 특허법 제42조의 규정에 의한 출원, 특허법 제60조의 규정에 의한 출원심사 를 청구합니다. 대리인
권태복 (인) 대리인
이화익 (인)
【수수료】

【기본출원료】	20	면	29,000 원
【가산출원료】	8	면	8,000 원
【우선권주장료】	0	건	0 원
【심사청구료】	14	항	557,000 원

【합계】	594,000 원
【감면사유】	정부출연연구기관
【감면후 수수료】	297,000 원
【기술이전】	
【기술양도】	희망
【실시권 허여】	희망
【기술지도】	희망
【첨부서류】	1. 요약서·명세서(도면)_1통

【요약서】

【요약】

본 발명은 유엠엘에스(UMLS : United Medical Language System)를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치 및 그 방법에 관한 것이다. 본 발명의 장치 및 방법은 유엠엘에스에서 메타시소러스를 제공받아 개체명 인식에 사용될 언어자원인 개념어, 단일어 및 범주키텀 데이터베이스를 각각 구축하고, 상기 개념어 데이터베이스에 저장된 각 개념어를 입력받아 상기 단일어 및 범주키텀 데이터베이스에 저장된 자료를 이용하여 각 개념어에 대한 자질을 추출하며, 상기 추출된 결과를 이용하여 개체명을 인식하기 위한 규칙 생성 및 규칙 필터링 과정을 거쳐 규칙 데이터베이스를 구축하며, 생물학 문서를 입력받아 개체명 후보가 되는 명사 및 명사구를 추출하여 상기 규칙 데이터베이스에 저장된 규칙을 상기 명사 및 명사구에 적용하여 개체명 인식을 수행한다. 이렇게 함으로써, 입력문서에서 중요 정보 개체로 활용될 수 있는 생물학적 개체명들을 효과적으로 추출할 수 있다.

【대표도】

도 1

【색인어】

생물학적 개체명, UMLS Metathesaurus, 자원 구축, 규칙 수집, 개체명 인식

【명세서】

【발명의 명칭】

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치 및 그 방법 {APPARATUS AND METHOD FOR RECOGNIZING BIOLOGICAL NAMED ENTITY FROM BIOLOGICAL LITERATURE BASED ON UMLS}

【도면의 간단한 설명】

도 1은 본 발명의 실시예에 따른 유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치의 전체 구성을 나타낸 도면.

도 2는 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법의 자원 구축 단계를 나타낸 도면.

도 3은 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법에서 개념어를 의미 범주에 따라 분할하기 위해 사용되는 MRCON 테이블과 MRSTY 테이블의 매핑 관계를 예시한 도면.

도 4는 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법의 규칙 수집 단계를 나타낸 도면.

도 5는 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법에서 생물학적 개체명의 특징을 반영하기 위해 정의된 자질들을 예시한 도면.

도 6은 상기 도 4에 도시된 자질 추출 단계의 보다 상세한 처리 흐름을 나타낸 도면.

도 7은 상기 도 4에 도시된 규칙 구성 단계의 보다 상세한 처리 흐름을 나타낸 도면.

도 8은 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법에서 사용되는 규칙의 표현예를 나타낸 도면.

도 9는 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법에서 특정 개념어를 대상으로 규칙을 구성한 예를 나타낸 도면.

도 10은 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법의 개체명 인식 단계를 나타낸 도면.

<도면의 주요부분에 대한 부호의 설명>

10 : 자원 구축부	11 : UMLS
12 : 개념어 D/B	13 : 단일어 D/B
14 : 범주키텀 D/B	20 : 규칙 수집부
21 : 규칙 D/B	30 : 개체명 인식부
31 : 문서 입력부	32 : 개체명 인식결과 출력부

【발명의 상세한 설명】

【발명의 목적】

【발명이 속하는 기술분야 및 그 분야의 종래기술】

<17> 본 발명은 유엠엘에스(UMLS : United Medical Language System, 이하 'UMLS'라 함)를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치 및 그 방법에 관한 것으로서, 생물학적 개체명을 인식하고 분류하는 기술에 관한 것이다.

<18> 생물학 연구의 활성화로 생물학 문헌이 급증하면서 이를 대상으로 한 높은 수준의 정보 추출 요구가 증대되고 있다. 단백질명, 유전자명, 실험 생물이나 생체 구성요소명 등은 중요한 생물학적 연구 결과를 기술하고 있는 생물학 문헌에서 정보의 핵심을 이루고 있다. 따라서, 생물학 문헌을 대상으로 한 정보추출을 수행하기 위해서는 이러한 생물학적 개체들의 이름을 정확하게 인식하고 분류하는 기술이 선행되어야 한다. 문헌을 대상으로 한 정보추출은 문헌 내의 정보 주체와 이들 간의 관계 또는 각 주체가 이끄는 정보의 흐름을 파악하는데 목적이 있다. 따라서, 생물학 문헌을 대상으로 한 정보추출의 경우에도 문헌 내의 정보주체가 되는 생물학적 개체명의 인식이 선행되어야 한다. 일반적으로, 생물학적 개체명을 인식하는 방법은 한정된 도메인을 대상으로 생물학적 지식을 갖춘 전문가가 대상 도메인에 대한 각종 언어 자원 및 규칙을 생성하고 이를 이용해 개체명을 인식하는 규칙 기반의 방식이 있다. 또한, 대용량의 생물학 문헌 학습 코퍼스(corpus)를 구축하고 이에 대해 기계학습 알고리즘을 적용해 개체명을 인식하는 통계 기반의 방식이 있다. 전자는 언어 자원 및 규칙 생성에 많이 비용이 소요되고, 후자는 생물학 문헌 학습 코퍼스 구축에 많은 비용이 소요된다는 문제가 있다.

<19> 선행 기술로서, 텍스트를 참조하여 새로운 이름들을 인식하고 추출하는 기술이 미국특허 제5,819,265호에 "Processing names in a text"이 1998년 10월 6일자로 등록되어 있다. 그러나, 상기 선행 특허는 UMLS 기반의 생물학 문헌을 처리하는 것에 대해서는 서술하지 않고 있으며, 문헌에 적은 빈도로 출현하는 이름이나 철자가 비슷하지만 의미가 다른 이름들이 출현할 경우에 오류 발생 가능성이 크다는 문제가 있다.

<20> 또 다른 선행 기술로서, David A. Campbell and Stephen B. Johnson에 의해 "A Technique for Semantic Classification of Unknown Words Using UMLS Resources"(Proceedings of American Medical Informations Association Symposium, pp 716-720)가 1999년 11월에 발표되

었고, Irena Spasic, Coran Nenadic and Sophia Ananiadou에 의해 "Using Domain Specific Verbs for Term Classification"(Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp 17-24)가 2003년 7월에 발표되었다. 상기 선행 논문들에 공개된 생물학적 개체명 인식 방법은 UMLS와 코퍼스를 동시에 이용하여야 하며, 패턴 규칙이 특정 형태로 제한되어 있어서 새롭게 생성되는 다양한 개체명을 인식하는 것에 한계가 있다.

【발명이 이루고자 하는 기술적 과제】

<21> 본 발명은 상기 설명한 종래의 기술적 과제를 해결하기 위한 것으로서, UMLS를 기반으로 생물학 문헌으로부터 정보의 주체가 되는 생물학적 개체명을 인식하는 장치 및 그 방법을 제공하는 것을 목적으로 한다.

<22> 보다 구체적으로, 본 발명은 생물학적 개체명의 특징을 반영한 다양한 자질과 UMLS라는 생물학 어휘 자원을 활용하여 개체명 인식 규칙을 구성하고, 이 규칙을 이용하여 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치 및 방법을 제공하는 것을 목적으로 한다.

【발명의 구성 및 작용】

<23> 상기한 목적을 달성하기 위한 본 발명에 따른 유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치는 유엠엘에스(UMLS : United Medical Language System)에서 메타시소러스를 제공받아 개체명 인식에 사용될 언어자원인 개념어, 단일어 및 범주키텀 데이터베이스를 각각 구축하는 자원 구축부; 상기 개념어 데이터베이스에 저장된 각 개념어를 입력받아 상기 단일어 및 범주키텀 데이터베이스에 저장된 자료를 이용하여 각 개념어에 대한 자

질을 추출하고, 상기 추출된 결과를 이용하여 개체명을 인식하기 위한 규칙 생성 및 규칙 필터링 과정을 거쳐 규칙 데이터베이스를 구축하는 규칙 수집부; 및, 생물학 문서를 입력받아 개체명 후보가 되는 명사 및 명사구를 추출하여 상기 규칙 데이터베이스에 저장된 규칙을 상기 명사 및 명사구에 적용하여 개체명 인식을 수행하는 개체명 인식부를 포함하는 것을 특징으로 한다.

<24> 또한, 상기 목적을 달성하기 위한 본 발명에 따른 유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법은, 유엠엘에스로부터 메타시소러스를 입력받아 개체명 인식을 위한 언어자원인 개념어, 단일어 및 범주키텀을 추출하여 각각에 대한 데이터베이스를 구축하는 자원 구축 단계; 상기 각 데이터베이스에 저장된 언어자원들을 이용하여 개념어의 자질을 추출하고 이에 대한 규칙을 구성하여 규칙 데이터베이스에 저장하는 규칙 수집 단계; 및, 문서를 입력받아 개체명 후보에 대한 자질을 추출하고, 상기 추출된 자질을 결합하여 개체명 후보를 결정하기 위한 규칙을 생성하며, 상기 규칙 데이터베이스에 저장되어 있는 규칙들과 상기 생성된 규칙을 매치하여 그 결과를 이용하여 최종 의미범주를 결정하는 개체명 인식 단계를 포함하는 것을 특징으로 한다.

<25> 이하, 본 발명의 바람직한 실시예에 따른 생물학적 개체명을 인식하는 장치 및 방법을 첨부한 도면을 참조하여 상세히 설명한다.

<26> 도 1에는 본 발명의 실시예에 따른 UMLS를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치의 전체 구성이 도시되어 있다.

<27> 상기 도 1에 도시된 바와 같이, 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 장치는 자원 구축부(10), 규칙 수집부(20) 및 개체명 인식부(30)를 포함한다. 아래에서 상기 각부의 동작을 보다 상세하게 설명한다.

<28> 상기 자원 구축부(10)는 UMLS(11)에서 메타시소러스(metathesaurus)를 제공받아 개체명 인식에 사용될 언어자원인 개념어(concept name), 단일어(single name) 및 범주키텀(category keyterm)에 대한 데이터베이스(12, 13, 14)를 각각 구축한다. 상기 규칙 수집부(20)는 상기 개념어 데이터베이스(12)에 저장된 각 개념어를 입력받아 상기 단일어 및 범주키텀 데이터베이스에 저장된 자료를 이용하여 각 개념어에 대한 자질을 추출하고, 상기 추출된 결과를 이용하여 개체명을 인식하기 위한 규칙 생성 및 규칙 필터링 과정을 거쳐 규칙 데이터베이스(21)를 구축한다. 상기 개체명 인식부(30)는 문서 입력부(31)를 통해 입력 문서 중 개체명 후보가 되는 명사 및 명사구를 추출하여 상기 규칙 데이터베이스(31)에 저장된 규칙을 상기 명사 및 명사구에 적용함으로써 개체명 인식을 수행한다.

<29> 다음으로, 도 2 내지 도 10을 참조하여 본 발명의 실시예에 따른 UMLS를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법에 대해 설명한다. 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법은 자원 구축 단계, 규칙 수집 단계 및 개체명 인식 단계로 이루어진다. 아래에서 상기 각 단계에 대해 도면을 참조하여 보다 상세하게 설명한다.

<30> 먼저, 도 2를 참조하여 자원 구축 단계에 대해 설명한다. 상기 도 2에는 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법의 자원 구축 단계가 도시되어 있다.

<31> 앞서 설명한 바와 같이, 상기 도 1의 자원 구축부(10)는 UMLS(11)에서 제공하는 메타시소러스를 이용하여 개체명 인식에 사용될 언어자원인 개념어 데이터베이스(12), 단일어 데이터베이스(13) 및 범주키텀 데이터베이스(14)를 구축한다. 상기 UMLS(11)로부터 상기 자원 구축부(10)에 제공되는 메타시소러스는 생의학 분야에서 사용되는 다양한 통제어휘들 및 분류 등에서 한 번 이상 나타난 개념들에 대한 정보를 포함하고 있다.

<32> 상기 도 2에서 자원 구축 단계가 시작되면, 가장 먼저 개념 집합 분할 단계(S100)가 수행된다. 상기 개념 집합 분할 단계(S100)에서는 상기 UMLS(11)의 메타시소러스에 포함된 테이블 중에서 개념어를 나타내는 각 문자열의 의미를 기술하기 위한 테이블인 MRCON과 각 개념어에 할당된 의미범주(semantic category)를 기술하기 위한 테이블인 MRSTY 테이블에 있는 정보를 도 3에 도시된 바와 같은 매핑 조건(mapping condition)을 이용하여 매핑하여 MRCON 테이블에 있는 데이터를 각 의미범주 별로 분할한다. 상기 도 3에는 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법에서 개념어를 의미 범주에 따라 분할하기 위해 사용되는 MRCON 테이블과 MRSTY 테이블의 매핑 관계가 예시되어 있다. 상기 도 3에 도시된 매핑 조건은 MRCON 테이블의 CUI(unique identifier for concept)와 MRSTY 테이블의 CUI가 일치하는 경우에 MRCON 테이블에 있는 데이터들 중 LAT(language of term)의 값이 "ENG"인 데이터들만을 MRSTY 테이블의 TUI(unique identifier of semantic type)에 해당하는 값에 따라 서로 다른 집합으로 분할하기 위한 조건이다.

<33> 다음으로, 개념어 추출 단계(S101)가 수행된다. 상기 개념어 추출 단계(S101)에서는 MRCON 테이블의 데이터를 각 의미범주 별로 분할한 개념집합 분할 단계(S100)의 결과로부터 MRCON 테이블의 STR(String) 필드에 있는 값들인 개념어들이 추출되어 개념어 데이터베이스(12)에 저장된다.

<34> 다음으로, 단일어 추출 단계(S102)와 범주키텀 추출 단계(S103)가 각각 수행된다. 상기 단일어 추출 단계(S102)에서는 상기 개념어 데이터베이스(12)로부터 단일 단어 자체가 개체명으로 사용되는 단어인 단일어들이 추출되어 단일어 데이터베이스(13)에 저장된다. 단일어의 경우, 여러 의미범주에서 사용될 수 있기 때문에 각 단일어가 사용되고 있는 의미범주에 대한 정보도 단일어 데이터베이스(13)에 함께 저장된다. 상기 범주키텀 추출 단계(S103)에서는 상

기 개념어 데이터베이스(12)로부터 특정 범주에서 주로 출현하여 개체명을 구성하는데 있어 중요한 역할을 하는 단어인 범주키텀이 추출되어 범주키텀 데이터베이스(14)에 저장된다. 상기 범주키텀은 개체명을 구성하는 각 단어가 가장 많이 출현한 의미범주에서의 분포([최다 출현 범주에서의 출현 빈도]/[모든 범주에서의 출현 빈도])를 계산하여 임계치로 필터링함으로써 얻어진다. 상기 단일어 추출 단계(S102)와 범주키텀 추출 단계(S103)가 완료되면, 상기 자원 구축 단계가 종료된다.

<35> 다음으로, 도 4를 참조하여 상기 규칙 수집 단계에 대해 설명한다. 상기 도 4에는 본 발명의 실시예에 따른 생물학적 개체명을 인식하는 방법의 규칙 수집 단계가 도시되어 있다.

<36> 상기 도 1의 규칙 수집부(20)는 개념어 데이터베이스(12)에 저장되어 있는 각 개념어를 입력받아 각 개념어를 구성하는 토큰(token)의 자질을 추출하며, 이와 같이 추출된 자질을 규칙의 형태로 구성하여 규칙 데이터베이스(21)를 구축한다. 상기 도 4에 도시된 규칙 수집 단계는 상기 규칙 수집부(20)에서의 동작을 나타내며, 상기 규칙 수집 단계는 자질 추출 단계(S200)와 규칙 구성 단계(S201)로 구성된다. 또한, 상기 자질 추출 단계(S200)와 규칙 구성 단계(S201)의 보다 구체적인 구성은 도 6 및 도 7에 각각 도시되어 있다.

<37> 도 4의 규칙 수집 단계가 시작되면, 가장 먼저 자질 추출 단계(S200)가 수행된다. 상기 자질 추출 단계(S200)에서는 생물학적 개체명의 특징을 반영하기 위해 정의된 도 5에 도시된 바와 같은 다양한 자질들을 이용하여 도 6에 도시된 바와 같은 자질 추출 흐름에 따라 각 개념어에 대한 자질을 추출한다. 생물학적 개체명들은 명명법상 대문자, 숫자 혹은 알파벳이 아닌 문자들을 포함하는 경우가 많기 때

문에 도 5의 (가)에 예시된 대문자 표현, 영숫자, 특수문자와 같은 자질들이 사용된다. 또한, 상기 생물학적 개체명들은 전치사나 접속사를 포함하는 경우와 개체의 기능 또는 범주를 나타내는 단어를 포함하는 경우가 있기 때문에 도 5의 (가)에 명시된 바와 같이 각각 전치사 또는 접속사 자질과 단일어 및 범주키텀 자질이 사용된다. 그리고, 개체명 토큰이 상기한 어떠한 자질에도 속하지 않는 경우를 나타내기 위해 도 5의 (가)에 예시된 기타라는 자질이 사용된다. 각 자질은 서브타입(subtype)을 가진다. 단일어나 범주키텀 자질은 개체명 인식을 위해 정의된 의미범주들을 자질의 서브타입으로 가지며, 대문자 표현, 영숫자, 전치사접속사, 특수문자, 그리고 기타 자질의 경우는 각각 도 5의 (나), (다), (라), (마), (바)에 명시된 바와 같은 서브타입을 가진다.

<38> 상기 도 4의 자질 추출 단계(S200)는 도 6에 보다 상세하게 도시되어 있으며, 이러한 자질 추출 단계(S200)는 도 4 및 도 10에서 각각 이용되며, 도 1의 규칙 수집부(20)와 개체명 인식부(30)에서 각각 개념어 및 개체명 후보를 입력받아 이들을 구성하는 각 토큰의 자질들을 추출하는데 사용된다. 다시 상기 도 6을 참조하여 자질 추출 단계(S200)의 동작을 보다 구체적으로 설명한다. 동작이 시작되면, 토큰화 단계(S2000)가 수행되며, 상기 토큰화 단계(S2000)에서는 개념어 데이터베이스(12)에 저장되어 있는 개념어들과 문서 입력부(31)에서 추출된 개체명 후보들이 공백문자 및 특수문자들을 이용해 토큰으로 분할된다. 상기 단계(S2000)에서 분할된 각 토큰들은 특수문자 인식 단계(S2001), 영숫자 인식 단계(S2002), 단일문자 인식 단계(S2003), 전치사 또는 접속사 인식 단계(S2004), 단위(unit) 인식

단계(S2005), 그리스어 인식 단계(S2006), 대문자 표현 인식 단계(S2007), 단일어 인식 단계(S2008), 범주키텀 인식 단계(S2009)를 순차적으로 거침으로써 도 5에 도시된 바와 같은 자질들이 해당 단계에서 추출된다. 상기 단일어 인식 단계(S2008) 및 범주키텀 인식 단계(S2009)에서는 토큰이 자원 구축부(10)에서 구축된 단일어 데이터베이스(13) 및 범주키텀 데이터베이스(14)에 존재하는지 검색하여 해당 단일어나 범주키텀의 서브타입이 얻어진다. 상기 범주 키텀 인식 단계(S2009)가 완료되면 상기 자질 추출 단계(S200)가 종료하며, 도 7의 규칙 생성 단계(S2010)로 점프하여 도 4에 도시된 규칙 구성 단계(S201)의 세부적인 절차가 수행된다.

<39> 상기 규칙 구성 단계(S201)는 도 7에 도시된 바와 같이 규칙 생성 단계(S2010)와 규칙 필터링 단계(S2011)로 구성된다. 상기 규칙 구성 단계(S201)에서는 상기 도 4의 자질 추출 단계(S200)에서 개념어를 토큰화하여 추출된 자질을 입력받아 개체명 인식을 위한 규칙이 생성되고, 생성된 규칙이 필터링되어 최종적으로 규칙 데이터베이스(21)가 구축된다.

<40> 상기 도 7을 참조하면, 규칙 생성 단계(S2010)에서는 도 8에 도시된 규칙 표현 방식과 같이 개념어의 각 토큰이 결합되어 개체명 인식을 위한 규칙이 생성된다. 이 때, 토큰이 단일어 자질을 갖는 경우, 하나의 단일어가 여러 개의 서브타입을 가질 수 있기 때문에 이러한 서브타입을 모두 고려한 규칙들이 생성되어야 한다. 도 9는 "Gas bacillus"라는 특정 개념어를 대상으로 규칙을 구성한 예를 도시한 도면이다. 상기 "Gas bacillus"는 "Gas" 및 "bacillus"라는 두 개의 토큰으로 분할된다. 각 토큰에서 자질을 추출하면, 상기 "Gas"는 대문자 표현 자질과 단일어 자질을 가지며, 상기 "bacillus"는 단일어 자질만을 가진다. "Gas" 및 "bacillus"의 단일어 자질이 각각 2개의 서브타입을 가지기 때문에 이들의 조합을 모두 고려하면, 도 9에 도시된 바와 같이, 4개의 규칙들이 만들어진다.

<41> 상기 도 7의 규칙 필터링 단계(S2011)에서는 상기 규칙 생성 단계(S2010)에서 만들어진 모든 규칙을 대상으로 [특정 범주에서의 규칙 출현 빈도]/[모든 범주에서의 규칙 출현 빈도]가 계산되고, 임계치로 필터링되어 규칙 데이터베이스(21)가 구축된다. 상기 규칙 필터링 단계(S2011)가 완료되면, 상기 규칙 구성 단계(S201)가 모두 종료된다.

<42> 상기와 같이 규칙 구성 단계(S201)가 종료되면, 상기 도 1에 도시된 개체명 인식부(30)에서 개체명 인식 단계가 수행되며, 도 10에는 상기 개체명 인식 단계의 보다 구체적인 구성이 도시되어 있다. 상기 개체명 인식부(30)는 문서 입력부(31)에서 제공되는 문서를 대상으로 상기 규칙 데이터베이스(21)의 규칙을 적용하여 개체명 인식을 수행한다. 아래에서는 도 10을 참조하여 개체명 인식 단계에 대해 보다 구체적으로 설명한다.

<43> 도 10에서 개체명 인식 단계가 시작되면, 가장 먼저 개체명 후보 추출 단계(S300)가 수행된다. 상기 개체명 후보 추출 단계(S300)에서는 문서 입력부(31)에서 제공되는 문서에 대해 형태소 분석이 수행되며, 개체명 후보가 되는 명사 및 명사구가 추출된다. 이 때, 명사구는 단순히 연속된 명사로 이루어진 구(phrase)만이 아니라 형용사, 관사, 전치사 및 접속사를 포함하는 형태의 명사구를 의미한다. 상기 개체명 후보 추출 단계(S300)가 완료되면, 자질 추출 단계(S200)가 수행된다. 상기 자질 추출 단계(S200)에서의 자질 추출은 상기 개체명 후보 추출 단계(S300)의 적용 결과로 얻어진 명사 및 명사구를 대상으로 수행된다. 이 때, 자질 추출 단계(S200)에서는 도 1의 자원 구축부(10)에서 구축된 단일어 데이터베이스(13) 및 범주키텀 데이터베이스(14)에 저장되어 있는 정보들을 이용하여 단일어 및 범주키텀 자질이 추출된다. 다음으로, 규칙 생성 단계(S2010)에서는 개체명 후보에 대해 자질 추출 단계(S200)를 적용하여 얻어지는 토큰들을 도 8에 도시된 규칙 표현 방식에 의해 결합함으로써 규칙이 생성된다.

<44> 다음으로, 규칙 매치 단계(S301)에서는 상기 규칙 생성 단계(S2010)에서 만들어진 개체명 후보에 대한 규칙과 상기 규칙 데이터베이스(21)에 저장되어 있는 규칙들을 완전 매치(exact match), 부분 매치(partial match) 또는 내포 매치(nested match)의 방식으로 매칭시킴으로써 개체명 후보에 적합한 기존 규칙들이 추출된다. 상기 완전 매치는 두 규칙이 정확하게 동일하게 매치되는 것을 의미하고, 부분 매치는 두 규칙의 앞이나 뒤 또는 중간 부분이 매치되는 것을 의미하며, 내포 매치는 규칙 안에 또 다른 매치되는 규칙이 존재하는 것을 의미한다.

<45> 다음으로 개체명 범주 결정 단계(S302)가 수행되며, 상기 단계(S302)에서는 규칙 매치 단계(S301)에서 추출된 기존 규칙들의 가중치와 개체명 범주 결정을 위한 몇 가지 휴리스틱(heuristic)을 이용하여 개체명 후보들의 최종적인 의미 범주가 결정되어 개체명 인식 결과 출력부(32)로 보내지며, 상기 개체명 인식 결과 출력부(32)에 의해 생물학적 개체명의 인식 결과가 제공된다.

<46> 상기 본 발명의 실시예에 따른 UMLS를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법은 컴퓨터 프로그램으로 제작되어서 하드디스크, 플로피 디스크, 광자기 디스크, 씨디 롬, 롬, 램 등의 기록매체에 저장될 수 있다.

【발명의 효과】

<47> 위에서 설명한 바와 같이, 본 발명에 따른 UMLS를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치 및 그 방법은 UMLS 메타시소러스를 이용하여 생물학적 언어 자원을 자동으로 구축하고, 이를 활용하여 생물학적 문헌에서 사용되는 개체명들을 자동으로 인식함

으로써 생물학적 개체명 인식 시스템을 구축하는데 필요한 노력 및 비용이 절감될 수 있다. 또한 생물학적 개체명 인식기를 전문가의 도움 없이도 도메인에 관계없이 빠르게 구현할 수 있어 생물학적 문헌을 대상으로 한 정보추출 연구 활성화에 기여할 수 있다.

<48> 이상에서 설명한 것은 본 발명에 따른 UMLS를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치 및 그 방법을 실시하기 위한 하나의 실시예에 불과한 것으로서, 본 발명은 상기한 실시예에 한정되지 않고, 이하의 특허청구 범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 분야에서의 통상의 지식을 가진 자라면 누구든지 다양한 변경 실시가 가능한 범위까지 본 발명의 기술적 정신이 있다고 할 것이다.

【특허청구범위】

【청구항 1】

유엠엘에스(UMLS : United Medical Language System)에서 메타시소러스를 제공받아 개체명 인식에 사용될 언어자원인 개념어, 단일어 및 범주키텀 데이터베이스를 각각 구축하는 자원 구축부;

상기 개념어 데이터베이스에 저장된 각 개념어를 입력받아 상기 단일어 및 범주키텀 데이터베이스에 저장된 자료를 이용하여 각 개념어에 대한 자질을 추출하고, 상기 추출된 결과를 이용하여 개체명을 인식하기 위한 규칙 생성 및 규칙 필터링 과정을 거쳐 규칙 데이터베이스를 구축하는 규칙 수집부; 및

생물학 문서를 입력받아 개체명 후보가 되는 명사 및 명사구를 추출하여 상기 규칙 데이터베이스에 저장된 규칙을 상기 명사 및 명사구에 적용하여 개체명 인식을 수행하는 개체명 인식부를 포함하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치.

【청구항 2】

제1항에 있어서,

상기 자원 구축부는 상기 유엠엘에스의 메타시소러스를 의미범주 별로 분할한 결과로부터 개념어를 추출하여 개념어 데이터베이스를 구성하고, 상기 개념어 데이터베이스에 저장되어 있는 개념어를 처리하여 단일어와 범주키텀을 추출하며, 상기 추출된 단일어와 범주키텀을 이용하여 단일어 데이터베이스와 범주키텀 데이터베이스를 각각 구축하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치.

【청구항 3】

제1항에 있어서,

상기 규칙 수집부는 상기 개념어 데이터베이스에 저장되어 있는 각 개념어를 구성하는 토큰의 자질을 추출하고, 그 결과를 결합하여 규칙을 생성하며, 상기 규칙에 가중치를 부여하여 임계치로 필터링하여 얻어진 결과를 상기 규칙 데이터베이스에 저장하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치.

【청구항 4】

제1항에 있어서,

상기 개체명 인식부는 상기 문서 입력부를 통해 제공되는 문서를 대상으로 개체명 후보를 추출하고, 개체명 후보를 구성하는 각 토큰의 자질을 추출한 결과를 결합하여 개체명 후보를 결정하기 위한 규칙을 생성하고, 상기 생성된 규칙을 규칙 데이터베이스에 저장되어 있는 규칙들과 매치하여 개체명 후보에 적합한 기존 규칙들을 추출하며, 추출된 각 규칙들의 가중치와 개체명 범주 결정을 위한 휴리스틱을 적용하여 개체명 후보에 대한 최종적인 의미범주를 결정하여 개체명을 인식하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 장치.

【청구항 5】

유엠엘에스로부터 메타시소러스를 입력받아 개체명 인식을 위한 언어자원인 개념어, 단일어 및 범주키텀을 추출하여 각각에 대한 데이터베이스를 구축하는 자원 구축 단계;

상기 각 데이터베이스에 저장된 언어자원들을 이용하여 개념어의 자질을 추출하고 이에 대한 규칙을 구성하여 규칙 데이터베이스에 저장하는 규칙 수집 단계; 및,

문서를 입력받아 개체명 후보에 대한 자질을 추출하고, 상기 추출된 자질을 결합하여 개체명 후보를 결정하기 위한 규칙을 생성하며, 상기 규칙 데이터베이스에 저장되어 있는 규칙들과 상기 생성된 규칙을 매치하여 그 결과를 이용하여 최종 의미범주를 결정하는 개체명 인식 단계를 포함하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 6】

제5항에 있어서,

상기 자원 구축 단계는

상기 메타시소러스에 포함된 테이블 중에서 개념어를 나타내는 각 문자열의 의미를 기술하기 위한 MRCON 테이블과 각 개념어에 할당된 의미범주를 기술하기 위한 MRSTY 테이블에 있는 정보를 매핑 조건을 이용하여 매핑함으로써 상기 MRCON 테이블에 저장되어 있는 데이터를 각 의미범주 별로 분할하는 제1단계;

개념집합 분할의 결과로부터 MRCON 테이블의 STR 필드에 있는 값들을 추출해 개념어 데이터베이스에 저장하는 제2단계;

개념어 데이터베이스로부터 단일어를 추출하여 단일어 데이터베이스에 저장하는 제3단계; 및

개념어 데이터베이스로부터 범주키텀을 추출하여 범주키텀 데이터베이스에 저장하는 제4단계를 포함하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 7】

제6항에 있어서,

상기 MRCON 및 MRSTY 테이블에 있는 정보를 매핑하는 조건은 MRCON 테이블의 CUI와 MRSTY 테이블의 CUI가 일치하는 경우 MRCON 테이블에 있는 데이터들 중 LAT 필드의 값이 "ENG"인 데이터들만을 MRSTY 테이블의 TUI에 해당하는 값에 따라 서로 다른 집합으로 분할하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 8】

제6항에 있어서,

상기 제4단계는 개념어 데이터베이스에 저장되어 있는 개념어들을 이용하여 개체명을 구성하는 각 단어가 가장 많이 출현한 의미범주에서의 분포를 계산하여 임계치로 필터링 하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 9】

제5항에 있어서,

상기 규칙 수집 단계는

개념어 데이터베이스에 저장되어 있는 각 개념어에 대해 토큰 별로 자질을 추출하는 제1 단계; 및,

자질이 추출된 토큰들을 결합해 규칙을 구성하고 이에 대해 가중치를 계산하여 필터링 한 결과를 규칙 데이터베이스에 저장하는 제2단계를 포함하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 10】

제9항에 있어서,

상기 제1단계는 생물학적 개체명의 특징을 반영하기 위해 정의된 자질들인 대문자 표현, 영숫자, 특수문자, 전치사 또는 접속사, 단일어 및 범주키텀 자질과 각 자질의 서브타입을 이용하여 개념어 데이터베이스에 저장되어 있는 각 개념어의 토큰들에 대한 자질을 추출하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 11】

제9항에 있어서,

상기 제2단계는

상기 제1단계에서 개념어를 토큰화하여 자질을 추출한 결과를 입력받아 토큰이 갖는 자질들의 서브타입에 따라 서브타입의 조합을 모두 고려한 개수만큼의 규칙들을 생성하는 단계; 및

상기 생성된 모든 규칙을 대상으로 각 범주에서 규칙의 출현 분포를 계산하고 임계치로 필터링하여 규칙 데이터베이스를 구성하는 규칙 필터링 단계를 포함하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 12】

제5항에 있어서,

상기 개체명 인식 단계는

입력 문서를 대상으로 개체명 후보가 되는 명사 및 명사구를 추출하는 개체명 후보 추출 단계;

개체명 후보의 각 토큰에 대해 자질을 추출하는 자질 추출 단계;

개체명 후보의 각 토큰에 대해 자질을 추출한 결과들을 결합하여 개체명 후보 결정을 위한 규칙을 생성하는 규칙 생성 단계;

생성된 규칙을 규칙 데이터베이스에 저장되어 있는 규칙들과 매치하는 규칙 매치 단계; 및,

개체명 후보들의 최종적인 의미범주를 결정하는 개체명 범주 결정 단계를 포함하는 것을 특징으로 하는

유엠엘에스를 기반으로 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 13】

제12항에 있어서,

상기 규칙 매치 단계는 개체명 후보를 결정하기 위한 규칙과 상기 규칙 데이터베이스에 저장되어 있는 규칙들을 완전 매치, 부분 매치 또는 내포 매치의 방식으로 매치하여 개체명 후보를 결정하기에 적합한 기존 규칙들을 추출하는 것을 특징으로 하는

유엠엘에스를 기반으로 하는 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【청구항 14】

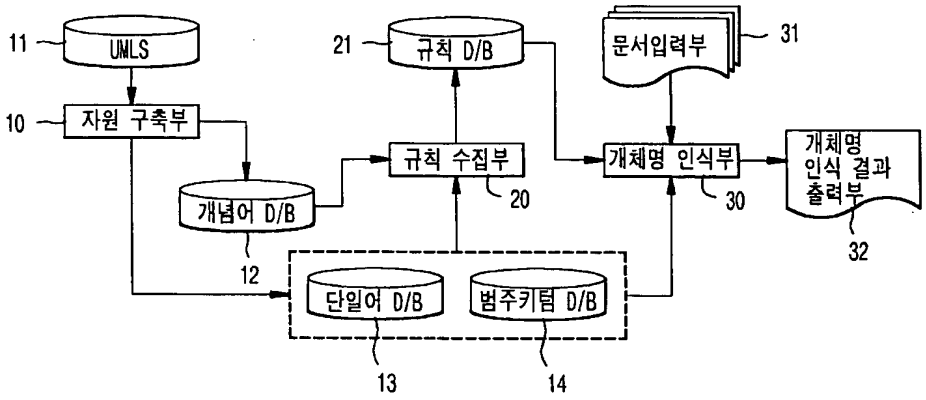
제12항에 있어서,

상기 개체명 범주 결정 단계는 규칙 매치 단계에서 추출된 기존 규칙들의 가중치와 개체명 범주 결정을 위한 휴리스틱을 이용하여 개체명 후보들의 최종적인 의미범주를 결정하여 개체명 인식 결과로서 출력하는 것을 특징으로 하는

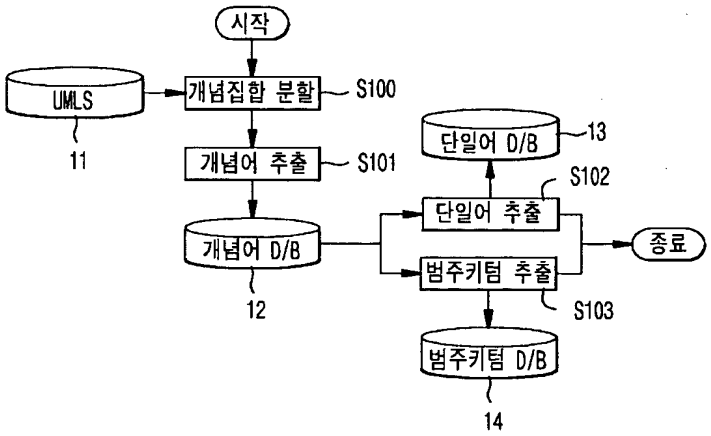
유엠엘에스를 기반으로 하는 생물학 문헌으로부터 생물학적 개체명을 인식하는 방법.

【도면】

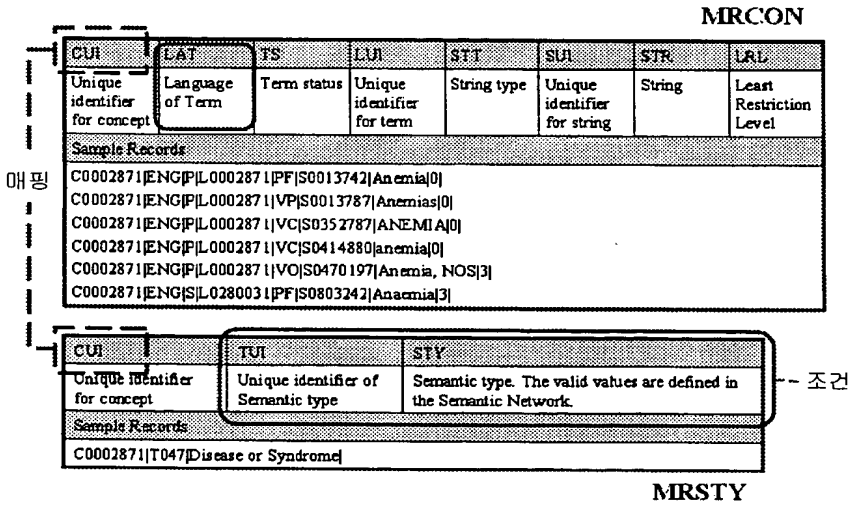
【도 1】


11
UMLS
10
자원 구축부
개념어 D/B
12
21
규칙 D/B
규칙 수집부
20
문서입력부
31
개체명 인식부
30
개체명 인식 결과 출력부
32
단일어 D/B
13
범주키텀 D/B
14

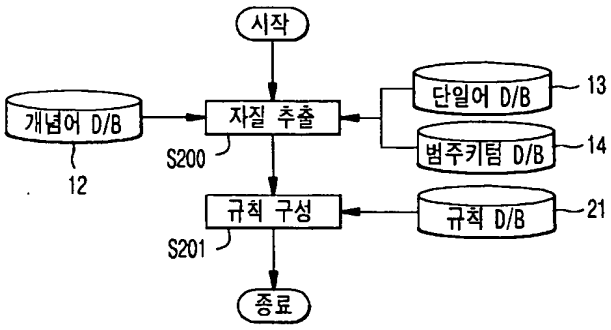
【도 2】


시작
UMLS
11
개념집합 분할
S100
개념어 추출
S101
개념어 D/B
12
단일어 D/B
13
S102
단일어 추출
범주키텀 추출
S103
종료
범주키텀 D/B
14

【도 3】


MRCON
CUI
LAT
TS
LUI
STT
SUI
STR
LRL
Unique identifier for concept
Language of Term
Term status
Unique identifier for term
String type
Unique identifier for string
String
Least Restriction Level
Sample Records
C0002871|ENG|P|L0002871|PF|S0013742|Anemia|0|
C0002871|ENG|P|L0002871|VP|S0013787|Anemias|0|
C0002871|ENG|P|L0002871|VC|S0352787|ANEMIA|0|
C0002871|ENG|P|L0002871|VC|S0414880|anemia|0|
C0002871|ENG|P|L0002871|VO|S0470197|Anemia, NOS|3|
C0002871|ENG|S|L0280031|PF|S0803242|Anaemia|3|
매핑
CUI
TUI
STY
Unique identifier for concept
Unique identifier of Semantic type
Semantic type. The valid values are defined in the Semantic Network.
조건
Sample Records
C0002871|T047|Disease or Syndrome|
MRSTY

【도 4】



【도 5】

(가)

자질	의미
단일어	단독으로 쓰여 개체명을 나타내는 단어
범주키텀	특정 범주에서 주로 사용되는 단어
대문자 표현	토큰을 이루는 문자열의 대소문자 구성 특징
영숫자	숫자를 포함하는 토큰
전치사·접속사	전치사 및 접속사
특수문자	특수문자
기타	그 밖의 경우

(나)

서브타입	예
START	Abstract
END	abstractA
ALL	ABCDE
MIXED	AbcDE, abAcde
ROMNUM	I, II, III, IV, ...

(다)

서브타입	예
ONLY	1234
YEAR	1900~2999
UNIT	10bp
GREEK	1alpha, alpha10
OTHER	10abc, abc10, ab10cd

(라)

about, above, across, after, afterward, against, al, along, alongside, also, amid, among, an, and, any, apart, around, as, at, athwart, bar, because, before, behind, below, beneath, beside, besides, between, beyond, but, by, concerning, considering, despite, down, downward, during, either, et, etc, except, for, forward, from, if, in, including, inside, into, inward, like, many, minus, more, much, near, neither, next, nor, nos, not, of, off, on, only, onto, onward, opposite, or, other, out, outside, outward, over, past, pending, per, plus, respecting, round, save, since, some, sp, such, that, the, through, throughout, till, to, too, toward, under, underneath, until, unto, up, upon, upward, versus, via, well, whether, with, within, without

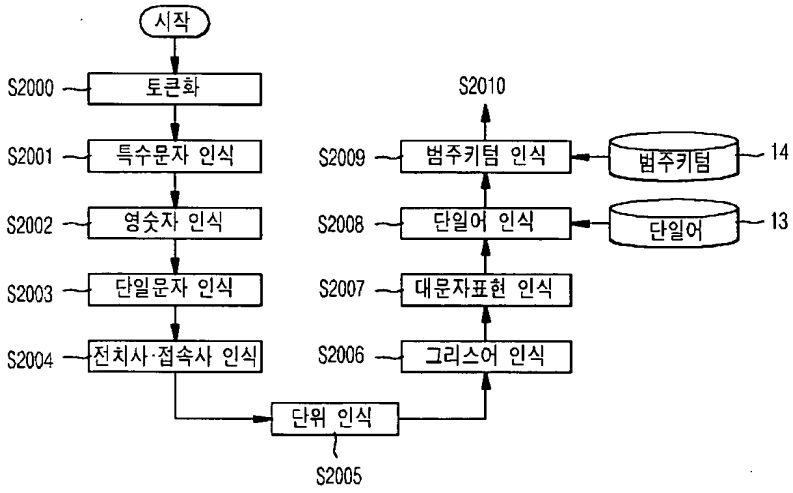
(마)

TAB, LF, VT, FF, CR, SPACE
!"#$%&'()*+,-./:;⇔?@[\]^_`{|}~

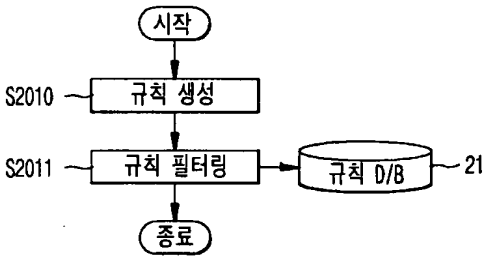
(바)

서브타입	예
UPA	A
LOWA	a
ALPHA	b-z, B-Z
OTHER	선행된 자질들 중 어느 곳에도 속하지 못하는 경우

【도 6】



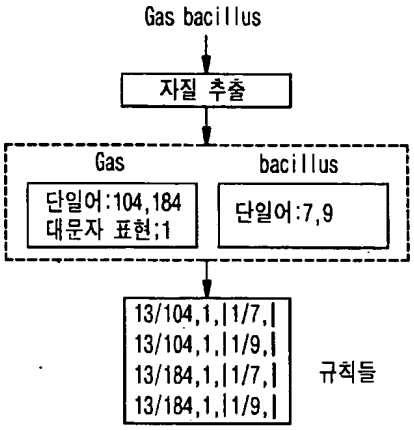
【도 7】



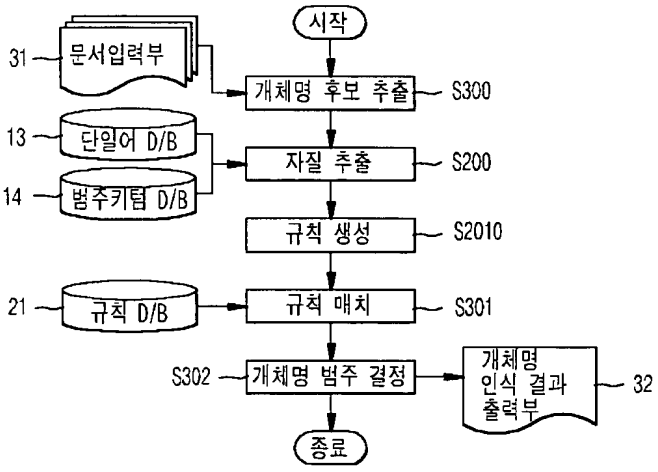
【도 8】

```
RULE := TOKEN | (TOKEN)*
TOKEN := FEATNUM / TYPENUM, (TYPENUM)*
FEATNUM := One of the (1, 2, 3, 4, 5, 6, 7, 12, 13, 23, 123)
                /*  1 = 단일어,    2 = 범주키텀,        3 = 대문자 표현,                */
                /*  4 = 영숫자,    5 = 전치사·접속사,   6 = 특수문자,      7 = 기타  */
                /*  12, 13, 23, and 123 are the combinations of the feature 1, 2, and 3.  */
TYPENUM := A Number that represents each of the feature's subtype.
```

【도 9】


Gas bacillus
자질 추출
Gas
bacillus
단일어:104,184
대문자 표현;1
단일어:7,9
13/104,1,|1/7,|
13/104,1,|1/9,|
13/184,1,|1/7,|
13/184,1,|1/9,|
규칙들

【도 10】


시작
31
문서입력부
개체명 후보 추출
S300
13
단일어 D/B
자질 추출
S200
14
범주키텀 D/B
규칙 생성
S2010
21
규칙 D/B
규칙 매치
S301
S302
개체명 범주 결정
개체명
인식 결과
출력부
32
종료